Scanning the Database for Recombinant HIV-1 Genomes

Adam C. Siepel and Bette T. Korber

MS K710, Los Alamos National Laboratory, Los Alamos NM 87545

INTRODUCTION

The recent identification of recombination events among sequences of different subtypes as an important source of new variation in HIV-1 [1] leads us to present the results of a comprehensive search for recombinant genomes in our HIV-1 master alignments. This volume of *Human Retroviruses and AIDS* includes results describing the *gag* and *env* coding regions, the two largest sequence sets within the database; a later release will include results obtained while searching the *pol*, *nef*, and *ltr* regions, for which less data is available. By publishing the results of a comprehensive search for recombinant genomes we hope to: (1) provide a better sense of the prevalence of intersubtype recombinant genomes, (2) enlarge the set of sequences available to researchers studying recombinant genomes, and (3) alert the HIV research community to "problem" sequences that should be avoided, for example, when looking for subtype "controls" with which a new sequence is to be compared.

Due to the large number of sequences in the database we decided to undertake a comprehensive search for recombinants only after developing a computer program capable of rapidly scanning individual sequences and providing summary information for sets of sequences [2]. At last count there were some 284 sequences in the HIV-1 *env* master alignment alone; such a large data set is prohibitive to the bootscanning approach [3] which requires the construction of numerous phylogenetic trees, and is therefore labor-intensive for both the computer and the researcher. The guiding principle behind the computer program we have developed, which we call the Recombinant Identification Program (RIP), is to pull out all sequences that show any signs of intersubtype mosaicism, so that they can be subjected to more sophisticated phylogenetic analysis. Thus the final steps remain similar to those utilized in the bootscanning approach, but the vast majority of sequences which show no signs of mosaicism are "weeded out", rendering much smaller the set to be painstakingly evaluated by hand.

RIP works by sliding a window of user-specified size along the length of a query sequence and evaluating, with each new position of the window, which subtype the query most resembles within its boundaries. It can be applied to nucleotide or amino-acid sequences. A query sequence is identified as a possible intersubtype recombinant if it bears a significant resemblance to one subtype in one window and another subtype in another window. Subtypes are described by consensus sequences, and the degree of relatedness between the query sequence and any particular subtype is defined in terms of the number of nonidentical bases (Hamming distance) between the query and that subtype's consensus sequence. The best match within each window is qualified by a measure of confidence obtained by comparing the distance to the best-matching subtype to the distance to the second-best-matching subtype. Confidence is calculated using a z-test, assuming: (i) that each site evolves independently according to the same process; and (ii) the binomial distribution that theoretically results from the use of Hamming distances can be approximated by a normal distribution. Note that measurements with respect to overlapping windows are not independent; for this reason and others the measure of confidence is approximate and is only used for heuristic purposes.

Each time RIP evaluates a query sequence, it accepts as input the query and an alignment of background sequences. The query sequence must be in alignment with the background set. RIP can dynamically create consensus data from a background alignment of individual sequences, or it can use predefined consensus sequences. If it is computing consensus sequences, RIP will accept a consensus threshold, a minimum frequency of occurrence necessary for the most common character to qualify as a consensus character, and will eliminate sites at which such a threshold is not achieved. The intent of using a consensus threshold is to eliminate highly variable sites which are likely to be homoplastic.

RIP is capable of processing a large set of query sequences at once, evaluating each one with the same parameter settings and against the same background set. After it is applied to a set of query sequences, the program provides summary tables describing possible recombinants and unlikely recombinants.

Several parameters are capable of changing RIP's sensitivity. Probably the most important of these is the size of the sliding window, which is specified as a number of base-pairs. Large windows give a clearer signal and can detect significance in long sections of weak similarity. Very small windows, on the other hand, are able to detect short mosaic segments, but tend to produce more noise and indicate significance where it cannot be proven. Another important parameter is a flag that switches *ON* or *OFF* the "informative mode". When the informative mode is active RIP counts only those mismatches which occur at positions where at least one subtype consensus sequence differs from the others.[1] Informative mode can improve the program's signal, particularly in well-conserved regions of an alignment. Other input parameters include a minimum level of certainty (as assessed by the method described above) required for a match to be labeled significant, the consensus threshold, and a flag that determines how the program handles gaps previously inserted to maintain alignment. The consensus threshold is used by the program as it calculates a consensus sequence: at a particular position, characters which do not appear with a frequency at least as great as the consensus threshold are left undefined (they are represented by question-marks). Mismatches that occur at positions at which any subtype consensus is undefined are not counted. Gaps can be handled in one of two ways: gap-stripping or gap-squeezing. Gap-stripping eliminates from consideration all columns in which the query sequence, or at least one subtype consensus sequence, is represented by a gap. Gap-squeezing eliminates columns in which the query and *all* subtypes are represented by a gap. When scanning a particular data set, RIP will identify different sets of possible recombinants depending on its parameter settings. For this reason, it is important to scan a set numerous times with a variety of parameter settings.

METHODS

We systematically examined the HIV-1 *gag* and *env* master alignments for recombinants, starting by scanning them with RIP at various parameter settings, then performing follow-up analysis on sequences identified as possible recombinants.

Scanning with RIP

Each region was scanned with RIP ten times at ten widely varying parameter settings. Sequences needed only to be picked out as possible recombinants in one of the ten runs in order to qualify for further analysis. The parameter settings employed were based on past experience with the program,[2] and on a "tuning" procedure designed to optimize the consensus threshold for a given data set and window size.[3] We used window sizes of 100, 150, and 200 base-pairs,[4] consensus levels of 0, 50, and 60%, and certainty thresholds of 90 and 95%. In all cases we handled gaps by gap-stripping, and set informative mode *OFF*.

[1] Note that this is a special meaning of the term "informative", distinct from the one discussed later in this article in the context of phylogenetics.

[2] Work with nucleotide sequences, which have generally given better results than amino acid sequences, has shown that window sizes of 100, 150, and 200 base pairs are adequate for achieving a wide range of sensitivity. This work has also shown that, though informative mode is useful for generating detailed RIP output with improved clarity, it should not be used when scanning large sets; the reason is that informative mode tends to artificially boost the estimated significance of best matches, and result in the identification of numerous possible recombinants that turn out to be ambiguous. In addition, experience with the program has shown that a certainty threshold of 90% works well in most cases, but a higher level is necessary in conjunction with small windows in variable regions.

[3] We will not go into detail about the tuning method here, but will report that the clearest signal for nucleotide sequences at all window sizes, for *gag* and *env*, occurred at a consensus threshold near 0% or between 50 and 60%.

[4] A larger 250 b.p. window was also used in conjunction with a consensus threshold of 60%. The reason for this additional parameter setting is that a high consensus threshold can disable enough sites that larger window sizes are needed in order to acquire meaningful results.

Follow-up Analysis

Follow-up analysis for each possible recombinant included two required steps and an optional third step. The first step was close examination of the detailed RIP output corresponding to the sequence. The second was to estimate the location of the apparent crossover site, using the method described by Robertson, et. al [6]. If the sequence still appeared to be a possible recombinant after the first two steps, we proceded to the third step, which involved dividing the sequence at the estimated crossover points and constructing separate phylogenetic trees of the resulting regions. Each of these three steps is discussed in more detail in the following paragraphs.

Step one. Examining the detailed RIP output can provide useful information that cannot be extracted from the summary output. For example, one can tell if a short match occurred in a region that is well-conserved across subtypes. In conserved regions, RIP sometimes finds significant similarity to a certain subtype when there is not enough phylogenetic information to support such a claim. In addition the detailed output provides information about the absolute similarity to the best match, and about which sites were eliminated because of gap-stripping or because of a failure to achieve the consensus threshold. Figure 1 shows an example of the detailed output for an apparent recombinant, CAR4039.

Step two. Step two of the follow-up procedure, the estimation of crossover breakpoints, was accomplished using a computer program written to perform the optimization procedure described by Robertson et. al [6]. The procedure depends on a four sequence alignment including the apparent recombinant as sequence one, a representative of each of the two recombining subtypes as sequences two and three, and an outgroup as sequence four. In this case the subtype representatives were chosen from a set of control sequences (discussed below) on the basis of similarity to the recombinant;[5] the most similar control in each of the two separate regions of the recombinant served as a representative of its subtype. The outgroup in all cases was an O-group sequence, MVP5180. Any column of the four-sequence alignment having two representatives each of two distinct characters is an informative site, in the sense that it favors one of three evolutionary histories.[6] These informative sites can be used as a way of estimating the crossover point in the following way: for any hypothetical crossover position in the four-sequence alignment, one can count the number of columns to the left favoring a grouping of the query with sequence number two, the number of columns to the left favoring a grouping with sequence three, the number of columns to the right favoring a grouping with sequence two, and the number of columns to the right favoring a grouping with sequence three. Further, the likelihood that the observed distribution of sites favoring groupings with sequences two and three might occur randomly can be assessed using a χ^2 test with one degree of freedom. The most likely crossover site is the hypothetical crossover site at which the observed distribution is least likely to occur randomly (highest χ^2 value). In addition to estimating the crossover location to the nearest informative site, this method also provides additional information about the likelihood that recombination has occurred at all. [7]

Step three. If close examination of RIP output did not indicate that the identification of a sequence as a possible recombinant was merely an artefact stemming from sampling issues or the use of consensus sequences, and if the distribution of informative sites according to the method above did not indicate that it was unlikely that a sequence was a recombinant, then we constructed phylogenetic trees describing the apparently recombining segments. First we compiled a full-length alignment including the apparent recombinant and a number of background sequences representing other subtypes. These background sequences were drawn from a previously established set of controls which will be discussed below.

[5] The locations of the separate regions were estimated from the detailed RIP output, and Hamming distance was used as an inverse measure of similarity.

[6] For example, if at a particular site, the query sequence and a representative of the B subtype share the same nucleic acid, and the outgroup and a representative of the A subtype share a different nucleic acid, then that site favors an evolutionary history in which the query diverged from the B after the A and B diverged from one another, and thus implies that the query should be classified as a B. Similarly, other informative sites will favor classification of the query as an A, and still others will indicate that the A and B representatives resemble one another more than either resembles the query sequence.

[7] Consider TZ017 in Table-2 as an example. This sequence was identified as a possible recombinant, and groups relatively well with different subtypes in neighbor-joining trees. Yet the informative sites in its apparent "D" segment are no more in favor of a D lineage than an A lineage.

Next this alignment was split at the estimated crossover points,[8] into two or three separate alignments, depending on whether there appeared to be one or two crossover points. After dividing a full-length alignment we constructed a neighbor-joining tree for each segment, using the PHYLIP software package [7]. In all cases we gap-stripped alignment segments before building trees. The number of columns retained after gap-stripping is reported in the legend beneath each tree that is displayed in this report. We used the Kimura two-parameter distance method, with a transition-to-transversion ratio of 1.3 in *env* and 1.8 in *gag*.[9] After building simple neighbor-joining trees we ran bootstraps with 100 replicates. Bootstrap values are printed at important nodes for each tree displayed.

Determining Control Sequences

The control sets from which background sequences used in steps two and three were extracted contained only full-length sequences that did not themselves exhibit mosaic behavior. For *env* and *gag* each, we first constructed a subset of the master alignment by eliminating all partial sequences,[10] previously identified recombinants, and sequences that had not been scanned during preliminary work done last summer [2]. Then we gap-stripped this alignment, divided it into consecutive 300 base pair segments, and constructed neighbor-joining trees for each segment. None of the sequences appeared to be a solid member of one subtype in one segment and a solid member of another subtype in another segment, but there were some outliers that grouped loosely with more than one subtype;[11] these were eliminated, and the resulting alignment was the final control group. From the control group we picked three representatives[12] of each subtype, selecting if possible for diversity of geographical origin and phylogenetic position within a clade (i.e., we avoided choosing two nearly identical sequences). These sequences served as a core group, and were used as representatives of subtypes other than the ones involved in an apparent recombinant genome. When building trees we drew background sequences from both the core and control groups. For example, when examining an apparent A-D recombinant, we took representatives of the B, C, E, F, G, and H subtypes from the core group, and representatives of the A and D subtypes from the control group. Core and control alignments for each HIV-1 coding region will be made available via the Human Retroviruses and AIDS Database's World Wide Web site (`http://hiv-web.lanl.gov`) and ftp site (`atlas.lanl.gov`, see the directory pub/aids-db/pub/aids-db/ALIGN/CONTROLS).

RESULTS

Results are presented according to coding region, with *env* preceding *gag* because the *env* data set is significantly larger. Also, the *env* set contains likely recombinants that have not previously been identified, and the *gag* set does not.

[8] Note that, using the χ^2 optimization method, crossover points are estimated to occur at some point between two informative columns in a four-sequence alignment; we chose to divide the alignment at a point midway between those two columns.

[9] These values produce trees with the maximum likelihood, using DNAML [7].

[10] Partial-length sequences were eliminated because including them would cause numerous columns to be removed by the gap-stripping procedure.

[11] It should be noted that certain subtypes are poorly defined in various of these 300 b.p. segments. For example, in the fifth segment of *env*, the D clade is positioned as a subclade among the B sequences, and the G's are interspersed among the A's. Subtype differentiation is also poor in the seventh segment of *env*, in which the A, G, and E subtypes formed a sort of "super-clade". The B's and D's are very close in the second and fourth segments of *gag*, and form one large clade in the third segment. The A's do not form a well-defined clade in the first segment of *gag*, nor do the H's in the third or fifth segment. Recall that the 300 b.p. segments were defined *after* gap-stripping, and that there are totals of eight segments in *env* and five segments in *gag*.

[12] There are only two full-length G's in *gag* and *env* and no full-length H's in *env*. Our plan was to include the partial H's in the *env* analysis only if we needed to analyse an apparent H hybrid, which we did not have to do.

Results of scanning the *env* master alignment

Table-1 lists all twenty-five HIV-1 *env* sequences identified by RIP as possible intersubtype recombinants. The locus name of each sequence is followed by columns listing the largest window size at which it was detected, the lowest consensus threshold (at the same window size) at which it was detected, and the level of certainty required for a best-match to be called significant. The next eight columns correspond to the eight different subtypes of HIV-1. Each column contains two numbers in each row: the first is the number of windows (or the number of positions of the sliding window) in which the sequence matched the subtype's consensus sequence *with threshold certainty*, and the second (in parentheses) is the largest number of those windows that occurred contiguously. The last column of Table-1 relates to work done to follow up RIP. Six of the sequences detected (labeled with a "P" in the last column of Table-1) had previously been identified as likely recombinants and were not analysed further. The other nineteen were subjected to analysis as described in the METHODS section. The possibility that a sequence is recombinant is strongly supported in three cases (labeled "S") and weakly supported in three other cases (labeled "W"). Follow-up analysis was ambiguous in the remaining thirteen cases (labeled "A"). There follows a more detailed account of the nineteen possible recombinants.

Strong evidence for recombination in *env*

Follow-up analysis revealed strong evidence supporting that AR15 [8], DI2ACD [9], and TZ016 [10] are recombinant between subtypes in *env*. All three were first detected by RIP at window sizes of 200 b.p. or greater (see Table-1), and matched two different subtypes with certainty over long contiguous stretches (the "C" portion of DI2ACD is the shortest observed, at 57 250-b.p. windows). Figure 2, Figure 3, and Figure 4 show separate phylogenetic trees for the apparently recombining regions of each sequence. Note that both DI2ACD and TZ016 group with sequences of two different subtypes with bootstrap values of 100 out of 100. AR15 groups with sequences of the F subtype with a bootstrap value of 100 in its Figure 2a, and with sequences of the B subtype with a bootstrap value of 92 in Figure 2b. The proximity of the B and D subtypes in Figure 2b may account for the slightly lower bootstrap values observed for both groups.

Table-2 shows the crossover points for sequences AR15, DI2ACD, and TZ016, as estimated by the method described earlier in this article. In each case the distribution of informative sites is strongly in favor of one subtype in one region of the sequence and another subtype in the other region.

In all three cases, the subtypes that appear to have recombined are consistent with what has been observed in the country of origin. AR15 originates from Argentina where B and F subtypes cocirculate [8], and TZ016 from Tanzania where A and D subtypes cocirculate [10]. DI2ACD is from Burundi; not much data from Burundi is available, but C and D subtypes are observed in nearby countries (see Table-7).

Weak evidence for recombination in *env*

Three more Tanzanian sequences from the same set as TZ016 [10] showed weak evidence for recombination in *env*: TZ005, TZ017, and TZ030. TZ005 and TZ030, which are closely related and have been classified as D's [10], both contain a 189-b.p. segment 162 b.p. downstream from their 5′ termini that groups with sequences of the C subtype. RIP caught TZ005 with a 150 b.p. window size and TZ030 with a 100 b.p. window size; in both cases a relatively small number of contiguous windows significantly matched the C subtype: 10 in the case of TZ005 and 32 in the case of TZ030. RIP also detected TZ017 at a window size of 150 b.p. Like TZ016, TZ017 groups with A's when its full length is considered, and also like TZ016, it has a stretch near its 5′ terminus that tends to group with the D's; in the case of TZ017, however, this stretch ends 265 b.p. sooner and, as will be discussed below, has a more tenuous similarity to D sequences.

Figure 5 shows phylogenetic trees including segments one, two and three of sequences TZ005 and TZ030. The grouping of TZ005 and TZ030 with the C's in segment two and with the D's in segment three is clear; in segment one, however, the tree is muddled slightly by the appearance of the B's as a subgroup within a larger B/D clade. The grouping of B's and D's together is not unusual, as was noted in the METHODS sections, especially in a segment as short as this one (144 sites retained, after gap-stripping). Because segment one contains limited information, and segment three shows a convincing grouping of TZ005 and TZ030 with their apparent subtype (bootstrap 100), we will focus here on segment two. The bootstrap value in segment two for the C clade, which includes TZ005 and TZ030, is only 41. Such a low value may result from the short length of the segment (203 sites retained) and its consequently limited amount of phylogenetic information; nevertheless it weakens any claim that these sequences are recombinants. Similarly, the crossover-point analysis of both TZ005 and TZ030 (see Table-2) revealed 15 informative sites in segment two, 10 of which favored grouping the sequences with C's rather than D's. All in all, it appears that TZ005 and TZ030 are notably similar to the C subtype in segment two, but that similarity is not strong enough to claim that it must be the result of a recombination event.

The "D portion" of TZ017, which occurs from bases 1 to 440 (see Table-2), is a similar case to the "C portions" of TZ005 and TZ030. Segment one of TZ017 groups with the D's with a bootstrap value of only 76. The informative sites in the crossover-point analysis were less convincing: as many sites favored grouping segment one with the A's (12 sites) as with the D's (background included SF1703 as an A and UG024 as a D). Again, TZ017 is similar to the D's in segment one, but that similarity need not have resulted from recombination.

Ambiguous evidence for recombination in *env*

Thirteen possible recombinants appeared ambiguous on further inspection. Nine were apparent A-G hybrids and one was an apparent A-E hybrid; these will be discussed in the context of a special relationship among the A, E, and G subtypes (see METHODS) that becomes apparent when small segments of the genome are analysed independently. The other three possible recombinants will be discussed individually.

The A, G, and E subtypes form a kind of "super clade" over some regions of the HIV-1 genome, indicating that the E's and G's could possibly have diverged from the A's, which has been suggested to be older. However it arose, the proximity of these subtypes over stretches of 250 base pairs or less can lead RIP to find similarity with threshold certainty to the G consensus, for example, in a relatively solid A subtype sequence. This problem is exacerbated by the extreme diversity within the A subtype [11] which allows diminished similarity between individual sequences and the consensus sequence. Here we see one of the pitfalls of comparing a query sequence to consensus sequences rather than to individual sequences in the database.

The 3′ terminus of GP41 is one particular region in which the A, E, and G subtypes are close [12]. It was in this region that RIP found significant similarity to the G consensus for sequences DJ258A, DJ264A, VI191A, and Z321, and significant similarity to the E consensus for sequence DJ263A. Figure 6 shows two separate trees corresponding to the two regions of *env* resulting from a division at the crossover point held roughly in common by the four apparent A-G's. Sequence DJ263A is included as well because it is closely related to DJ258A and DJ264A. In Figure 6a, although the the five sequences in question are outliers to the core group formed by the other six A's, the A clade is generally well-defined. In Figure 6b, however, the A, E, and G clades are closer, and the five A outliers are nearer to the G subtype than to the A subtype. It is not hard to imagine finding windows within this region in which these sequences match the G consensus more closely than the A consensus. The match in DJ263A to the E consensus is thought to result from a window size of only 100 b.p., and from the the fact that the A's, E's, and G's are particularly jumbled in the region identified, as can be seen in a neighbor-joining tree of that short region. It is notable that the crossover-point analysis, in the cases of DJ264A, VI191A and Z321, revealed a significant preference for a grouping in segment two with the G subtype, and in the case of DJ263A, for a grouping with the E subtype. However Figure 6 clearly indicates that DJ258A, DJ263A, DJ264A, VI191A, and Z321 are not classifiable in the segment in question. The "mosaic" character of these five sequences could be as readily explained by differential rates of evolution as by recombination.

Similar issues surround the five sequences from the Central African Republic indicated by RIP as A-G recombinants: CAR286A, CAR4023, CAR4054, CAR4081, and CAR423A. These sequences, which have all been classified as A's except for CAR4081 (unclassified), resemble the G's over a lengthy segment at the $5'$ end. Figure 7 shows this segment along with its complement. Here the grouping with the G sequences is perhaps more solid than in the case discussed above, but the apparent hybrids are still outliers to both the A and G subtypes. Also the crossover-point analysis shows that, in four out of the five sequences, a grouping with the A's is actually favored in the apparent G segment. Again, from the evidence here, it seems likely that these five CAR sequences are not recombinants.

The other three sequences identified by RIP with ambiguous follow-up results are 91US006.10, ELI, and VI354. 91US006.10, which has also been called USHOBR-10, showed significant similarity to the D consensus in three contiguous windows at a window size of 100 b.p. Because the window size was small and the region that was identified is relatively well-conserved across subtypes (particularly the B and D subtypes), and because the crossover-point analysis revealed only 3 versus 2 informative sites favoring the D subtype within the region, there appears to be little evidence to support that 91US006.10 is a B-D recombinant. VI354 matched the A consensus in just three contiguous windows as well, though these were 250 b.p. windows. Again the region was conserved across subtypes, and the crossover-point analysis turned up unconvincing numbers in the region in question (9 versus 4 sites in favor of the A consensus). Oddly, VI354 which is classified as an F, showed nearly as few matches to the F consensus (five contiguous windows) as to the A consensus. This effect is thought to be the result of background sampling: the F consensus is weighted toward the South American F's (from Brazil and Argentina) which are rather different from the African F's (mostly from Cameroon), because the database contains more of the former and they are generally longer. ELI, classified as a D, significantly matched the A consensus in five contiguous windows. Again, the windows were small (100 b.p.), the region was relatively well conserved across subtypes, and the crossover-point analysis turned up only three informative sites in the apparent A region. There is not enough evidence to claim that ELI is an intersubtype recombinant.

Results of scanning the *gag* master alignment

Table-3 lists the fifteen HIV-1 *gag* sequences identified by RIP as possible intersubtype recombinants. The form is the same as that of Table-1, and again, discussion will be limited to previously unidentified sequences. In this case follow-up analysis revealed nothing more than ambiguous evidence indicating that any one of the nine previously unidentified sequences had resulted from a recombination event.

Ambiguous evidence for recombination in *gag*

All nine newly identified sequences appeared ambiguous on further inspection. These nine included five apparent A-G hybrids (one of which is an A-C-G), and single apparent B-D, D-F, D-H, and B-G-H hybrids. It appears that the A-G's, the D-H, and the B-G-H were picked out due to a lack of subtype robustness over two particular regions, similar to that observed in *env* between the A and G subtypes; these sequences will be discussed in groups according to the regions on the *gag* gene at which the subtype breakdowns occur. The B-D and D-F sequences will be discussed individually.

CI59, DJ258, and LBV2310 are closely related sequences that were picked out as possible A-G recombinants. All three sequences have been classified as A's [13], but RIP found 9-15 contiguous 100-b.p. windows near their midpoints that matched the G consensus with threshold certainty. χ^2 optimization estimated the apparent crossover points to be exactly the same on each sequence (see Table-4). The segment from bases 750 to 969 resembled the G representative rather than the A representative by 5 informative sites to 1 in CI59 and LBV2310, and 4 to 1 in DJ258. Note that there was a single match to the C subtype in DJ258, but that it appeared to be an artefact of a short window placed in a relatively well-conserved region, with a high consensus threshold making several sites unusable. Figure 8 shows three trees including sequences CI59, DJ258, and LBV2310, corresponding to the first A segment (segment one), the G segment (segment two), and the second A segment (segment three). Subtypes are reasonably well-defined in segments one and three, with the possible exception of the closely related B and D subtypes in segment one. In segment two however, the D, G, and H clades have broken down, and the cluster containing the three sequences in question as well as CI20, has moved away from the bulk of the A's and closer to the G's. This behavior is probably best explained by a lack of sufficient phylogenetic information in a relatively well-conserved segment that includes only 224 sites after gapstripping. There is not enough evidence here to justify a claim that CI59, DJ258, and LBV2310, or CI20 for that matter, are recombinant genomes.

Sequences SE365 and VI557, classified as a D and an H respectively, were picked out as possible recombinants based on similarity to alternate subtypes in a shorter region within the one described above. Crossover-point analysis revealed two informative sites favoring the alternate subtype in each case. By looking at segment two of Figure 8 we get an idea of what happens with SE365 and VI557 in the general region that was identified. Indeed VI557 is closer to the G's in segment two than it is to the other H, VI525. Also, SE365 has moved away from the other two D's and towards VI525, an H. There is no indication that these sequences are recombinants. Instead the behavior of all five of the sequences discussed above in segment two of Figure 8 serves as an example of how subtype analysis can become difficult when the region in question is short relative to its density of phylogenetic information.

Two other sequences classified as A's, K29 and LBV23, were identified as having a G region near their 5′ terminus. It appears that this 132-b.p. region is another one in which the distinction between A's and G's breaks down, particularly because the A's are more divergent here than in other segments of *gag* (see METHODS). RIP detected 20 and 28 contiguous G windows in this region, respectively, with a consensus threshold of 60% and window sizes of 100 and 200 b.p. The high consensus, which resulted in several question-marks in the G consensus, may have eliminated sites that actually favored the A subtype, and helped result in a significant match to the G's. The crossover-point analysis found only two and one informative sites, respectively, in segment one, possibly because the A's and G's are difficult to distinguish here. It found the preference for the G's to be null or negligible.

The remaining two sequences identified as possible recombinants in *gag* are WEAU160 and Z2Z6, and the follow-up analysis of both was unconvincing. In WEAU160 (classified as a B), RIP found 16 contiguous windows, spanning the region from position 460 to position 674, that matched the D subtype. As it turns out, this region is highly conserved across subtypes, and the B's and D's in particular are nearly indistinguishable. It appears that again, the sequence was identified only because of sampling. RIP found 8 contiguous windows in Z2Z6, which is classified as a D, that significantly matched the F consensus. Crossover-point analysis revealed a preference for the F's of 5 informative sites to 1 (see Table-4), but tree analysis showed Z2Z6 to be a D outlier in the region in question, relatively close to the F's, but not part of the F clade.

CONCLUSIONS

The Recombinant Identification Program (RIP) was applied to the HIV Sequence Database's master alignments for the *env* and *gag* coding regions. 284 *env* sequences and 91 *gag* sequences were scanned. RIP pulled out 25 possible recombinants from the *env* alignment, correctly matching the remaining 259 with their proper subtypes (see Table-5). Of the 25 possible recombinants, six had been previously identified as likely recombinants [1,2,12,14]. Follow-up analysis on the remaining nineteen strongly supported that three were recombinants, weakly supported that three others were recombinants, and was ambiguous with respect to the rest. The three new strongly-supported recombinants are sequences AR15 (B-F), DI2ACD (C-D), and TZ016 (A-D). Each sequence groups with more than one subtype with high bootstrap values, and appears, in a four-sequence alignment with representatives of the apparently combining subtypes, to be significantly more closely related to sequences of different subtypes in different regions. The three weakly-supported recombinants include sequences TZ005 (C-D), TZ017 (A-D), and TZ030 (C-D). These sequences showed a notable similarity to multiple subtypes, but that similarity need not have resulted from recombination. In the *gag* coding region RIP found 15 possible recombinants; again, nonmosaic sequences fell into the proper subtype categories. Of the 15 possible recombinants, six had been previously identified [1] and the other nine all appeared ambiguous on follow-up analysis.

RIP is capable of rapidly scanning large sets of data with reasonable accuracy, if numerous tests with varying parameter settings are performed. It picked out six of the seven intragenic recombinants identified by Robertson, et. al in *gag*, and three of the four that the same group identified in *env*. RIP did not identify sequence MAL in *env* as an A-D recombinant, or sequence VI354 in *gag* as an A-G. It appears to have missed MAL, which is a D over most of *env*, because its A-like region is only 98 b.p. long [1]. Most likely it did not pick out VI354 because of problems relating to the detection of A-G hybrids, which are discussed below.

Although it reduced the set of sequences to be examined closely to about one-tenth its original size, RIP still identifies sequences as possible recombinants, which on further inspection are not supported. Table-5 shows the number of sequences identified as possible recombinants in each coding region, which had been previously identified, were supported by strong evidence, were supported by weak evidence, or proved ambiguous. One can see that nine of fifteen of the *gag* sequences identified, and thirteen of twenty-five of the *env* sequences, turned out to be ambiguous. Many of these were identified because consensus sequences, which the program uses as representatives of subtypes, are inadequate in regions where a subtype's members are diverse. Other false positives occur due to sampling in regions that are well-conserved across subtypes. Still the analysis performed on these ambiguous sequences is not wasted. It is useful to know about possible hybrids, even if they cannot be proved.

Though the subtypes of HIV-1 are generally well-defined when full coding regions are evaluated, some subtypes are quite close in shorter regions. The A and G subtypes and B and D subtypes, in particular, are difficult to distinguish in many segments. Also, in at least one segment of *env*, the E subtype merges with the A and G subtypes. Such breakdowns in subtype differentiation, which become more frequent as the segments examined become shorter, can make assessing claims of A-G, A-E and B-D mosaicism difficult or even impossible. In fact RIP identified 5 possible A-G recominants in *gag* and nine in *env*, as well as single possible B-D recombinants in *env* and *gag* each; none of these turned out to be supported.

Table-6 and Table-7 summarize the likely HIV-1 intersubtype recombinants identified thus far. Table-6 lists them along with their estimated crossover breakpoints. Table-7 shows that many originate from countries in which the apparently recombining subtypes are co-circulating. Note that more than half of the fifteen hybrid genomes listed are A-D's from Central Africa or B-F's from South America.

The current version of RIP, written in C++ for UNIX, is available via the Human Retroviruses and AIDS Database's World Wide Web site (`http://hiv-web.lanl.gov`) and ftp site (`atlas.lanl.gov`, in the directory pub/aids-db/PROGS/RIP). The program is accompanied by limited documentation, which will be improved in early 1996. Note that RIP is not limited to use with subtypes, but can evaluate query sequences with respect to any clearly defined clades. We hope to develop the program further, as time allows, perhaps implementing a representation of sequence sets based on frequency of occurrence that would replace the simple consensus sequence, the use of nucleotide and amino acid distance matrices in similarity calculations, and an improved system for comparing regions with gaps.

ACKNOWLEDGMENTS

Special thanks are due to Wouter Janssens for allowing us to analyse and discuss sequence AR15 prior to its publication in *AIDS Research and Human Retroviruses*, to Beatrice Hahn for helpful conversations, and to Catherine Macken and Aaron Halpern for advice on statistical issues.

REFERENCES

[1] Robertson, D. L., Sharp, P. M., McCutchan, F. E., and B. H. Hahn. 1995. Recombination in HIV-1. *Nature* **374**:124–126.

[2] Siepel, A. C., Halpern, A. L., Macken, C., and B. Korber. 1995. A computer program designed to screen rapidly for HIV Type 1 intersubtype recombinant sequences. *AIDS Res Hum Retroviruses* **11**(11):1413–1416.

[3] McCutchan, F.E., Salminen, M., and D.S. Burke: Genotyping of HIV-1. In: *Human Retroviruses and AIDS 1995*. Los Alamos National Laboratory, Theoretical Biology and Biophysics, Los Alamos, New Mexico, 1995.

[4] Faulkner, D.V. and A. Jurka. 1988. Multiple aligned sequence editor (MASE). *Trends in Biochem Science* **13**:321–322.

[5] Korber, B. and G. Myers. 1992. Signature Pattern Analysis: A Method for Assessing Viral Sequence Relatedness. *AIDS Res. Hum. Retroviruses* **8**:1549–1559.

[6] Robertson, D. L., Hahn, B. H., and P. M. Sharp. 1995. Recombination in AIDS viruses. *J. Mol. Evol.* **40**:249–259.

[7] Felsenstein, J. 1989. PHYLIP - phylogeny inference package. *Cladistics* **5**:164–166.

[8] Campodonico, M., Janssens, W., Heyndrickx, L., Fransen, K., Leonaers, A., Fay, F.F., Taborda, M., van der Groen, G., and O.H. Fay. 1995. HIV-1 subtypes in Argentina and genetic heterogeneity of the V3 region. *AIDS Res Hum Retroviruses* (In press).

[9] Ranjbar, S., Slade, A., Jenkins, A., Heath, A., Kitchen, P., Almond, N, Osmanov, S. and H. Holmes. 1995. Molecular Characterization of an HIV Type 1 Isolate from Burundi. *AIDS Res Hum Retroviruses* **11**(8):981–984.

[10] Siwka, W., Schwinn, A., Baczko, K., Pardowitz, I., Mhalu, F., Shao, J., Rethwilm, A., and V. ter Meulen. 1994. *vpu* and *env* sequence variability of HIV-1 isolates from Tanzania. *AIDS Res Hum Retroviruses* **10**(12):1753–1754.

[11] Korber, B.T.M., Allen, E.E., Farmer, A.D., and G.L. Myers. 1995. Heterogeneity of HIV-1 and HIV-2. *AIDS* **9**(suppl A):S5–S18.

[12] Gao, F., Morrison, S.G., Thornton, C.L., Craig, S., Karlsson, G., Sodroski, J., Morgado, M., Galvao-Castro, B., von Briesen, H., Beddows, S., Weber, J., Robertson, D.L., Sharp, P.M., Shaw, G.M., Hahn, B.H., and the WHO and NIAID Networks for HIV Isolation and Characterization. 1995. Molecular cloning and analysis of functional envelope genes from HIV-1 sequence subtypes A through G. *J Virol*, submitted.

[13] Louwagie, J., McCutchan, F. E., Peeters, M., Brennan, T. P., Sanders-Buell, E., Eddy, G. A., van der Groen, G., Fransen, K., Gershy-Damet, G.-M., Deleys, R., and D. S. Burke. 1993. Phylogenetic analysis of *gag* genes from 70 international HIV-1 isolates provides evidence for multiple genotypes. *AIDS* **7**:769–780.

[14] Sabino, E. C., Shpaer, E. G., Morgado, M. G., Korber, B. T. M., Diaz, R. S., Bongertz, V., Cavalcante, S., Galvao-Castro, B., Mullins, J. I., and A. Mayer. 1994. Identification of human immunodeficiency virus type 1 envelope genes recombinant between subtypes B and F in two epidemiologically linked individuals from Brazil. *J. Virol.* **68**:6340–6346.

Table 1 Summary of Possible Recombinants Identified in the *env* Master Alignment

Locus	*Win*	*Con*	*Cert*	A	B	C	D	E	F	G	H	Follow-up
91US006.10	100	0	95	0(0)	883(705)	0(0)	3(3)	0(0)	0(0)	0(0)	0(0)	A
93BR019.10	200	50	90	0(0)	139(139)	0(0)	0(0)	0(0)	1730(1682)	0(0)	0(0)	P
AR15	200	50	90	0(0)	228(228)	0(0)	0(0)	0(0)	245(245)	0(0)	0(0)	S
CAR286A	100	50	95	99(97)	0(0)	0(0)	0(0)	0(0)	0(0)	25(25)	0(0)	A
CAR4023	150	60	90	91(91)	0(0)	0(0)	0(0)	0(0)	0(0)	163(163)	0(0)	A
CAR4039	200	60	90	121(121)	0(0)	0(0)	0(0)	646(364)	0(0)	0(0)	0(0)	P
CAR4054	100	50	95	111(75)	0(0)	0(0)	0(0)	0(0)	0(0)	29(29)	0(0)	A
CAR4081	200	0	90	415(415)	0(0)	0(0)	0(0)	0(0)	0(0)	132(132)	0(0)	A
CAR423A	200	50	90	280(280)	0(0)	0(0)	0(0)	0(0)	0(0)	205(205)	0(0)	A
DI2ACD	250	60	90	0(0)	0(0)	57(57)	594(594)	0(0)	0(0)	0(0)	0(0)	S
DJ258A	200	50	90	1639(1639)	0(0)	0(0)	0(0)	0(0)	0(0)	8(8)	0(0)	A
DJ263A	100	0	95	665(339)	0(0)	0(0)	0(0)	1(1)	0(0)	0(0)	0(0)	A
DJ264A	200	60	90	1412(1117)	0(0)	0(0)	0(0)	0(0)	0(0)	34(34)	0(0)	A
ELI	100	50	95	5(5)	0(0)	0(0)	321(112)	0(0)	0(0)	0(0)	0(0)	A
K124A	100	60	95	223(123)	0(0)	0(0)	344(342)	0(0)	0(0)	0(0)	0(0)	P
RJI01.5	200	50	90	0(0)	651(459)	0(0)	0(0)	0(0)	70(70)	0(0)	0(0)	P
TZ005	150	50	90	0(0)	0(0)	10(10)	425(425)	0(0)	0(0)	0(0)	0(0)	W
TZ016	250	60	90	156(156)	0(0)	0(0)	187(187)	0(0)	0(0)	0(0)	0(0)	S
TZ017	150	60	90	174(105)	0(0)	0(0)	105(105)	0(0)	0(0)	0(0)	0(0)	W
TZ030	100	60	95	0(0)	0(0)	32(32)	123(75)	0(0)	0(0)	0(0)	0(0)	W
UG266A	200	50	90	79(79)	0(0)	0(0)	1644(1644)	0(0)	0(0)	0(0)	0(0)	P
VI191A	250	60	90	1016(1016)	0(0)	0(0)	0(0)	0(0)	0(0)	6(6)	0(0)	A
VI354	200	60	90	3(3)	0(0)	0(0)	0(0)	0(0)	5(5)	0(0)	0(0)	A
Z321	100	0	95	284(252)	0(0)	0(0)	0(0)	0(0)	0(0)	2(2)	0(0)	A
ZAM184	200	50	90	45(44)	0(0)	189(189)	0(0)	0(0)	0(0)	0(0)	0(0)	P

Summary table showing all sequences identified by RIP as possible recombinants during one or more of the ten scans performed on the *env* coding region. The second, third, and fourth columns describe the parameter-settings at which sequences were identified: window size is reported in base-pairs (*Win*), and consensus (*Con*) and certainty (*Cert*) thresholds in percent. If a sequence was identified at several different parameter settings, as most were, the largest window size and lowest corresponding consensus threshold are listed. Columns five through twelve correspond to the eight subtypes. Two numbers are listed at the position at which each locus name intersects with each subtype letter. The first is the number of windows in which the RIP found significant similarity to a given subtype's consensus sequence when scanning a given query sequence. The second (in parentheses) is the largest number of those windows which occurred contiguously. Column thirteen describes the results of the follow-up analysis performed on each possible recombinant: we found strong (*S*), weak (*W*), or ambiguous (*A*) evidence that the sequence resulted from a recombination event. Sequences that had previously been identified as likely recombinants (*P*) were not analyzed further. Note that all results were obtained by scanning nucleotide sequences with gap-stripping mode operative.

Table 2 Estimated Crossover Locations on *env* Sequences Identified as Possible Recombinants

Locus	Background Sequences	Subtype	Region	Informative Sites 1	2	3
91US006.10	JRFL	B	1–1958	54	10	16
	JY1	D	1961–2087	2	3	0
		B	2147–2582	17	4	6
93BR019.10	BZ126A	F	1–197	5	0	1
	US3	B	227–275	0	6	1
		F	309–2246	82	15	10
		B	2307–2555	1	15	3
AR15	BZ126A	F	1–390	18	5	6
	JRFL	B	416–870	1	14	5
CAR286A	LBV217	G	1–745	12	13	6
	SF1703	A	746–1549	10	28	11
CAR4023	LBV217	G	1–744	11	18	8
	SF1703	A	745–1552	6	26	14
CAR4039	TH966	E	1–594	19	4	3
	KENYA	A	606–865	3	8	3
		E	867–1483	24	9	9
CAR4054	LBV217	G	1–571	11	13	10
	SF1703	A	619–1486	7	28	16
CAR4081	LBV217	G	1–697	20	12	5
	RW020	A	720–1486	10	22	15
CAR423A	LBV217	G	1–723	15	17	8
	RW020	A	724–1516	7	29	11
DI2ACD	SM145A	C	1–306	17	7	0
	UG274A	D	342–1458	10	42	10
DJ258A	KENYA	A	1–1886	73	15	24
	UG975	G	1991–2574	11	12	10
DJ263A	KENYA	A	1–2171	72	22	25
	TH022	E	2181–2242	0	5	0
		A	2328–2586	5	4	7
DJ264A	KENYA	A	1–1877	73	18	24
	UG975	G	1990–2562	7	13	11

Table 2 (cont.) Estimated Crossover Locations on *env* Sequences Identified as Possible Recombinants

Locus	Background Sequences	Subtype	Region	Informative Sites		
				1	2	3
ELI	JY1	D	1–348	12	2	3
	SF1703	A	414–508	0	3	4
		D	543–2558	71	25	25
RJI01.5	JRFL	B	1–412	18	4	3
	BZ163A	F	457–527	0	5	0
		B	529–1264	26	1	5
TZ005	UG024	D	1–153	8	1	1
	MW965	C	162–351	5	10	1
		D	369–1113	28	4	11
TZ016	UG024	D	1–705	17	11	7
	KENYA	A	765–1113	3	20	4
TZ017	UG024	D	1–440	12	12	1
	SF1703	A	463–1122	4	28	5
TZ030	UG024	D	1–153	8	1	1
	MW965	C	162–351	5	10	1
		D	369–1110	24	6	12
VI191A	KENYA	A	1–1890	63	15	23
	UG975	G	1914–2565	9	14	11
VI354	BZ163A	F	1–531	19	12	15
	KENYA	A	549–882	4	9	1
Z321	KENYA	A	1–1907	77	19	32
	UG975	G	1920–2568	9	18	15

Results of χ^2 analysis for possible recombinants identified while scanning the *env* coding region, presented in a form similar to the one used by Robertson, et. al [1]. Analysis was not performed on K124A, UG266A, or ZAM184 for which results have already been published [1] and are reproduced in Table-6. Each sequence was aligned with three others: a representative of each apparently recombining subtype and an outgroup (in all cases, MVP5180). Numbers of phylogenetically informative sites supporting each of three lineages are listed in the columns labeled *1*, *2*, and *3*. In lineage *1*, the apparent recombinant is most closely related to the representative of first subtype; in lineage *2*, it is most closely related to the representative of the second subtype; and in lineage *3* it is most closely related to the outgroup. The locus names of representatives of the first and second subtypes are shown in appropriate order in the column labeled *Background Sequences*. Crossover points were located so as to maximize statistical significance of the difference in the distribution of sites supporting phylogenies *1* and *2*, as assessed by a χ^2 test with one degree of freedom.

Table 3 Summary of Possible Recombinants Identified in the *gag* Master Alignment

Locus	Win	Con	Cert	A	B	C	D	F	G	H	Follow-up
BZ200	200	60	90	0(0)	632(632)	0(0)	0(0)	35(35)	0(0)	0(0)	P
CI32	150	60	90	278(149)	0(0)	0(0)	64(64)	0(0)	0(0)	0(0)	P
CI59	100	50	90	679(268)	0(0)	0(0)	0(0)	0(0)	9(9)	0(0)	A
DJ258	100	60	90	464(172)	0(0)	1(1)	0(0)	0(0)	15(15)	0(0)	A
G141	100	60	90	233(101)	0(0)	0(0)	28(28)	0(0)	0(0)	0(0)	P
K124	200	60	90	864(864)	0(0)	0(0)	248(248)	0(0)	0(0)	0(0)	P
K29	100	60	90	776(775)	0(0)	0(0)	0(0)	0(0)	20(20)	0(0)	A
LBV105	100	60	90	224(100)	0(0)	0(0)	0(0)	6(6)	20(20)	0(0)	P
LBV23	100	60	90	579(290)	0(0)	0(0)	0(0)	0(0)	28(28)	0(0)	A
LBV2310	100	50	90	772(524)	0(0)	0(0)	0(0)	0(0)	9(9)	0(0)	A
MAL	200	50	90	866(801)	0(0)	0(0)	62(62)	0(0)	0(0)	0(0)	P
SE365	100	60	90	0(0)	0(0)	0(0)	247(190)	0(0)	0(0)	15(15)	A
VI557	100	60	90	0(0)	2(2)	0(0)	0(0)	0(0)	15(15)	33(22)	A
WEAU160	200	60	90	0(0)	610(321)	0(0)	16(16)	0(0)	0(0)	0(0)	A
Z2Z6	100	60	90	0(0)	0(0)	0(0)	484(227)	8(8)	0(0)	0(0)	A

Summary table showing all sequences identified by RIP as possible recombinants during one or more of the ten scans performed on the *gag* coding region. The second, third, and fourth columns describe the parameter-settings at which sequences were identified: window size is reported in base-pairs (*Win*), and consensus (*Con*) and certainty (*Cert*) thresholds in percent. If a sequence was identified at several different parameter settings, as most were, the largest window size and lowest corresponding consensus threshold are listed. Columns five through eleven correspond to the seven subtypes represented in *gag*. Two numbers are listed at the position at which each locus name intersects with each subtype letter. The first is the number of windows in which RIP found significant similarity to a given subtype's consensus sequence when scanning a given query sequence. The second (in parentheses) is the largest number of those windows which occurred contiguously. Column twelve describes the results of the follow-up analysis performed on each possible recombinant. We found all hitherto unrecognized recombinants to be ambiguous (*A*). Sequences that had previously been identified as likely recombinants (*P*) were not analyzed further. Note that all results were obtained by scanning nucleotide sequences with gap-stripping mode operative.

Table 4 Estimated Crossover Locations on *gag* Sequences Identified as Possible Recombinants

Locus	Background	Subtype	Region	Informative Sites		
	Sequences			1	2	3
CI59	VI32	A	1–733	22	6	15
	LBV217	G	750–969	1	5	3
		A	972–1458	14	8	4
DJ258	VI32	A	1–733	24	4	14
	LBV217	G	750–969	1	4	2
		A	972–1458	14	10	3
K29	LBV217	G	1–132	1	1	0
	VI32	A	145–1467	13	47	14
LBV23	LBV217	G	1–132	1	0	1
	VI32	A	168–1470	12	42	21
LBV2310	VI32	A	1–733	23	4	11
	LBV217	G	750–969	1	5	3
		A	972–1458	14	8	5
SE365	UG274	D	1–780	23	1	6
	VI525	H	804–816	0	2	0
		D	910–1473	15	8	3
VI557	VI525	H	1–742	29	8	9
	LBV217	G	834–864	0	2	0
		H	885–1467	15	9	9
WEAU160	SF2	B	1–408	10	2	5
	UG270	D	475–475	0	1	0
		B	483–1503	28	3	11
Z2Z6	UG274	D	1–945	23	7	8
	BZ162	F	946–1074	1	5	3
		D	1080–1503	11	6	7

Results of χ^2 analysis for possible recombinants identified while scanning the *gag* coding region. Analysis was not performed on BZ200, CI32, G141, K124, LBV105, or MAL for which results of this form have already been published [1] and are reproduced in Table 6. Each sequence was aligned with three others: a representative of each apparently recombining subtype and an outgroup (in all cases, MVP5180). Numbers of phylogenetically informative sites supporting each of three lineages are presented in the columns labeled *1*, *2*, and *3*. In lineage *1*, the apparent recombinant is most closely related to the representative of first subtype; in lineage *2*, it is most closely related to the representative of the second subtype; and in lineage *3* it is most closely related to the outgroup. The locus names of representatives of the first and second subtypes are shown in appropriate order in the column labeled *Background Sequences*. Crossover points were located so as to maximize statistical significance, as assessed by a χ^2 test with one degree of freedom.

Table 5 Number of Sequences in Master Alignment Identified as Nonmosaic, by Subtype

Region	Subtype									
	A	B	C	D	E	F	G	H	O	U
gag	21*	28	7	9	0	4	4	1	2	0
env	29	114	25	23	18	12	13	2	5	1

Sequences identified as nonmosaic showed significant similarity to no more than one subtype consensus during any of the ten scans of each coding region.

Numbers of Sequences Identified as Possible Intersubtype Recombinants, by Subtype Pair†

	Evidence for	Apparently Recombining Subtypes											
Region	Recombination	A-C	A-D	A-E	A-G	A-F	B-D	B-F	B-H	C-D	D-F	G-H	Totals
gag	prev ident	0	4	0	1*	0	0	1	0	0	0	0	6
	strong	0	0	0	0	0	0	0	0	0	0	0	0
	weak	0	0	0	0	0	0	0	0	0	0	0	0
	ambiguous	0	0	0	5	0	1	0	1	0	1	1	9
env	prev ident	1	2‡	1	0	0	0	2	0	0	0	0	6
	strong	0	1	0	0	0	0	1	0	1	0	0	3
	weak	0	1	0	0	0	0	0	0	2	0	0	3
	ambiguous	0	1	1	9	1	1	0	0	0	0	0	13

* VI354 (*gag*) was identified as a nonmosaic A by RIP, even though it has been identified by Robertson, et al. [1] as an A-G recombinant.

† DJ258 (*gag*) was actually identified as having segments matching the A, C, and G subtypes, with a single C window; it was however, counted here as an A-G. Similarly, VI557 (gag) had B, G, and H segments but here the two contiguous B segment were ignored and it was counted as a G-H.

‡ Not including MAL, which was identified by Robertson, et al. [1] as having a 98 b.p. A segment in *env* but was not detected by RIP, presumably because this segment is so short.

Table 6 Summary of Likely HIV-1 Recombinants and Estimated Crossover Breakpoints

Locus name	Origin	Gene	Subtype	Region	Informative sites		
					1	2	3
K124	Kenya	*gag*	A	1-1050	35	3	4
		gag	D	1083–1477	0	14	3
		env	A	1–1065	30	4	4
		env	D	1096–2435	4	46	10
		env	A	2480–2580	7	1	1
MAL	Zaire	*gag*	A	1–1068	34	6	4
		gag	D	1101–1518	4	12	1
		env	D	1–2435	9	61	30
		env	A	2482–2580	5	1	0
UG266	Uganda	*env*	A	1–199	10	1	4
		env	D	252–2556	9	62	11
CI32	Cote d'Ivoire	*gag*	A	1–333	10	4	3
		gag	D	417–423	0	4	0
		gag	A	456–1477	21	7	10
G141	Gabon	*gag*	D	1–261	1	13	5
		gag	A	306–1459	38	7	5
VI354	Gabon	*gag*	A	1–1154	38	9	14
		gag	G	1245–1465	1	5	1
LBV105	Gabon	*gag*	A	1–1166	34	7	12
		gag	G	1209–1483	4	6	3
ZAM184	Zambia	*env*	C	1–328	0	11	0
		env	A	363–1053	13	2	12
		env	C	1068–1263	1	8	1
		env	A	1270–2547	33	7	13
BZ200	Brazil	*gag*	B	1–1227	38	1	6
		gag	F	1253–1474	1	11	2
RJI01	Brazil	*env*	B	1–412	18	4	3
		env	F	457–527	0	5	0
		env	B	529–1264	26	1	5
CAR4039	Cent Afr Rep	*env*	E	1–594	19	4	3
		env	A	606–865	3	8	3
		env	E	867–1483	24	9	9
93BR019	Brazil	*env*	F	1–197	5	0	1
		env	B	227–275	0	6	1
		env	F	309–2246	82	15	10
		env	B	2307–2555	1	15	3
AR15	Argentina	*env*	F	1–390	18	5	6
		env	B	416–870	1	14	5
DI2ACD	Burundi	*env*	C	1–306	17	7	0
		env	D	342–1458	10	42	10
TZ016	Tanzania	*env*	D	1–705	17	11	7
		env	A	765–1113	3	20	4

The table presented by Robertson, et. al [1] with newly identified sequences added and P-values left out (due to the fact that simulations were not performed for the added sequences). As in Table-2 and Table-4, numbers of phylogenetically informative sites supporting each of three lineages are listed in the columns labeled *1*, *2*, and *3*, and crossover points have been located so as to maximize statistical significance in the distribution of sites supporting phylogenies 1 and 2 (according a χ^2 test with one degree of freedom). Note that SIVcpz was used as an outgroup when determining informative sites for sequences K124 through BZ200 while MVP5180 was used as an outgroup for RJI01 through TZ016. RJI01 was identified as a likely recombinant by Sabino, et. al [14] and 93BR019 by Gao, et. al [12]. Region boundaries and distributions of informative sites reported for 93BR019 differ slightly from those reported by Gao, et. al [12], presumably due to the use of different background sequences.

Table 7 HIV-1 Subtype Cocirculation and Recombination

Country	Recombinants Observed	No. Sequences in V3 Master Alignment, HIV Sequence Database							
		A	B	C	D	E	F	G	H
Argentina	B-F	0	0	0	0	0	0	0	0
Brazil	B-F(3)	0	65	1	0	0	5	0	0
Burundi	C-D	0	0	0	0	0	0	0	0
Cenral Afr Rep	A-E	14	0	1	1	11	0	1	0
Cote d'Ivoire	A-D	21	0	0	1	0	0	0	0
Gabon	A-D, A-G(2)	1	1	0	0	0	1	3	0
Kenya	A-D	20	0	0	3	0	0	0	0
Tanzania	A-D	6	0	0	18	0	0	0	0
Uganda	A-D	26	0	2	53	0	0	1	0
Zambia	A-C	0	0	3	0	0	0	0	0
Zaire	A-D	14	0	0	9	0	0	0	1
S Amer countries*	B-F(4)	0	65	1	0	0	5	0	0
Cent Afr countries†	A-C, A-D, A-E, C-D(4),	80	0	6	84	11	0	2	1

The likely recombinants from Table-6, presented by country. The geographical distribution of subtypes demonstrates that many apparent recombinants originate in countries where the recombining subtypes are cocirculating. The numbers of sequences have been taken from the V3 master alignment, the largest available set of unique (nonclonal) sequences. Note that a country-by-country breakdown provides no ready explanation for the B-F recombinant from Argentina, the C-D from Burundi, or the A-C from Zambia; we can attempt to explain these, however, by combining the geographically close South American countries and Central African countries (see the last two lines of the table).

*Argentina and Brazil.
†Burundi, the Central African Republic, Kenya, Tanzania, Uganda, Zaire, and Zambia.

```
CAR4039    TAGTGATTAGATCTGAAAATATCACAAACAATGCCAAAACCATAATAGTACAGCTGGTTACGCCTGTACAAATTAATTGTACCAGACCC---TCCAACAATATAAGA---ACAAGT
A          ..??.................................................?.????.A.......A?..................T...A?........C........A....
B          ....A...............T....??.......T.....................AA?GAAT.....G.............A.........AA........C........A....
C          ..A.A...............C.G..........T......A...........T..TAA.?AAT.....G.....GTG.....A.........AA...T....C........A....
D          ..A.A...............C.......?.....T....?...............TAA.GA.T.....??............A..G.......A.?.?....?.......CA...?
E          ..A.A..C............C............................G..C..TAA..AAT.....G....C............................C.............
F          ..A.A..C......C........T..G.T.....A.................??.TAA.GAAT.......?...........A.........AA........C........A....
G          ..A.A..............C?.....G...........GT.........G.....TAA..AA?..?..G......??...............AAT.......C........A....
H          ..A....C..................G.....A......A...............TAAA.AT......GT......C...............AA...T..C.CG.......G....
BEST MATCH AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAEEEEEEEEEEEEEEEEEEE   EEEEEEEEEEEEEEE   EEEEEE
.          ++^^^^^^^^^                                                       EEEE                                ^^^^^   ^^^++^
```

Figure 1. Detailed RIP output describing the general region of the second of two apparent crossover sites in CAR4039. Users of the program can view output within the sequence editor MASE [4] or print it to hardcopy. The top line shown is the query sequence, the next eight lines represent signature patterns [5] of subtype consensus sequences with respect to the query sequence, and the bottom two lines indicate the subtypes that the query most resembles in various windows. In the signature patterns, periods (".") represent identity between the consensus and query sequences, letters represent consensus bases not present in the query sequence, and question marks ("?") indicate positions lacking a clearly defined consensus. Carets ("^") beneath the subtype names indicate a statistical confidence of 90% and plus signs ("+") indicate a confidence of 95%. The *BEST MATCH* for any given window is reported at the center of that window. If multiple subtypes match equally well, they are all reported. Subtype letters are shown in upper-case if the absolute similarity is above 90% and in lower-case if it is not. Here the crossover site can be approximated by the point at which A's give way to E's, which is between positions 874 and 879 (note in Table-2 that the same crossover site was estimated by a more sophisticated method to occur between positions 865 and 869). To produce this output, the window size was 200 base pairs, the consensus threshold was 60%, the certainty threshold was 90%, gaps were stripped, and informate mode was inactive.

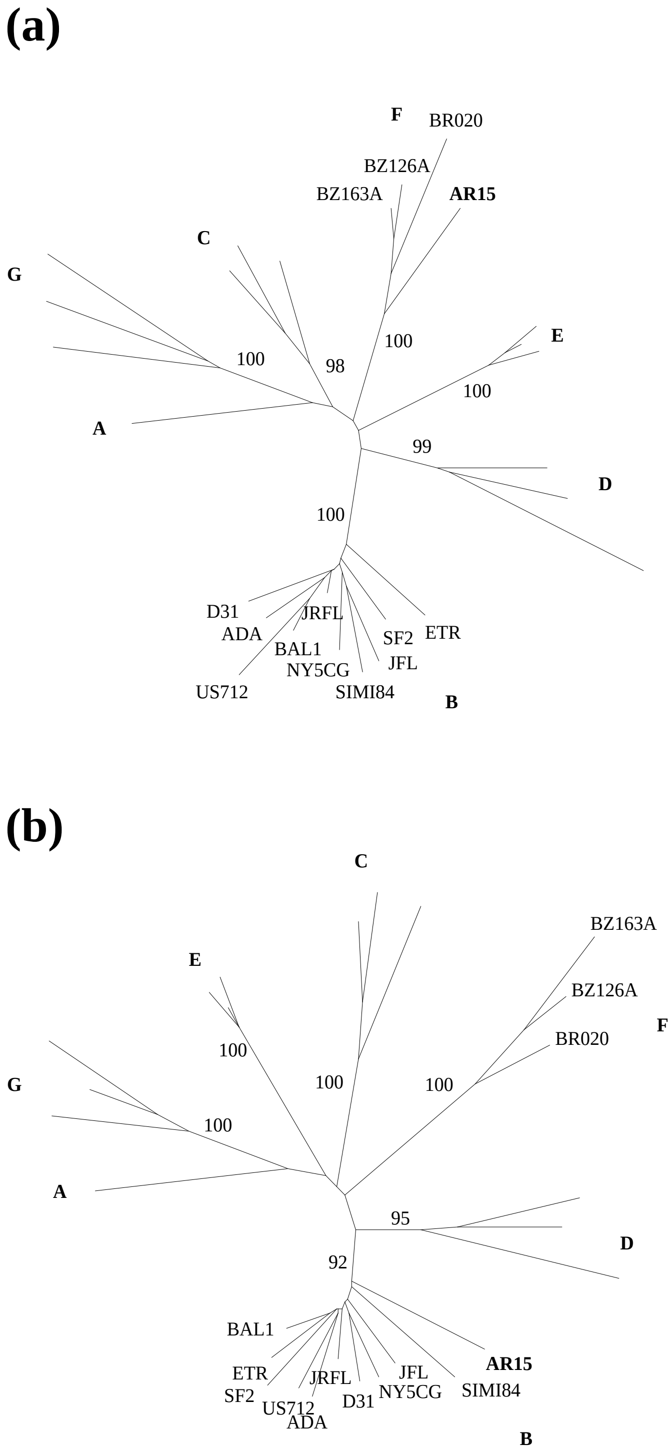

Figure 2. Phylogenetic trees showing two segments of sequence AR15, an apparent B-F recombinant in *env*. Note that AR15 groups with the B clade in (a) (base 1–403, 368 sites retained after gap-stripping) and with the F clade in (b) (bases 404–870, 447 sites retained after gap-stripping). All trees were generating with the neighbor-joining method using PHYLIP [7]. Important nodes are labeled with bootstrap values (100 replicates).

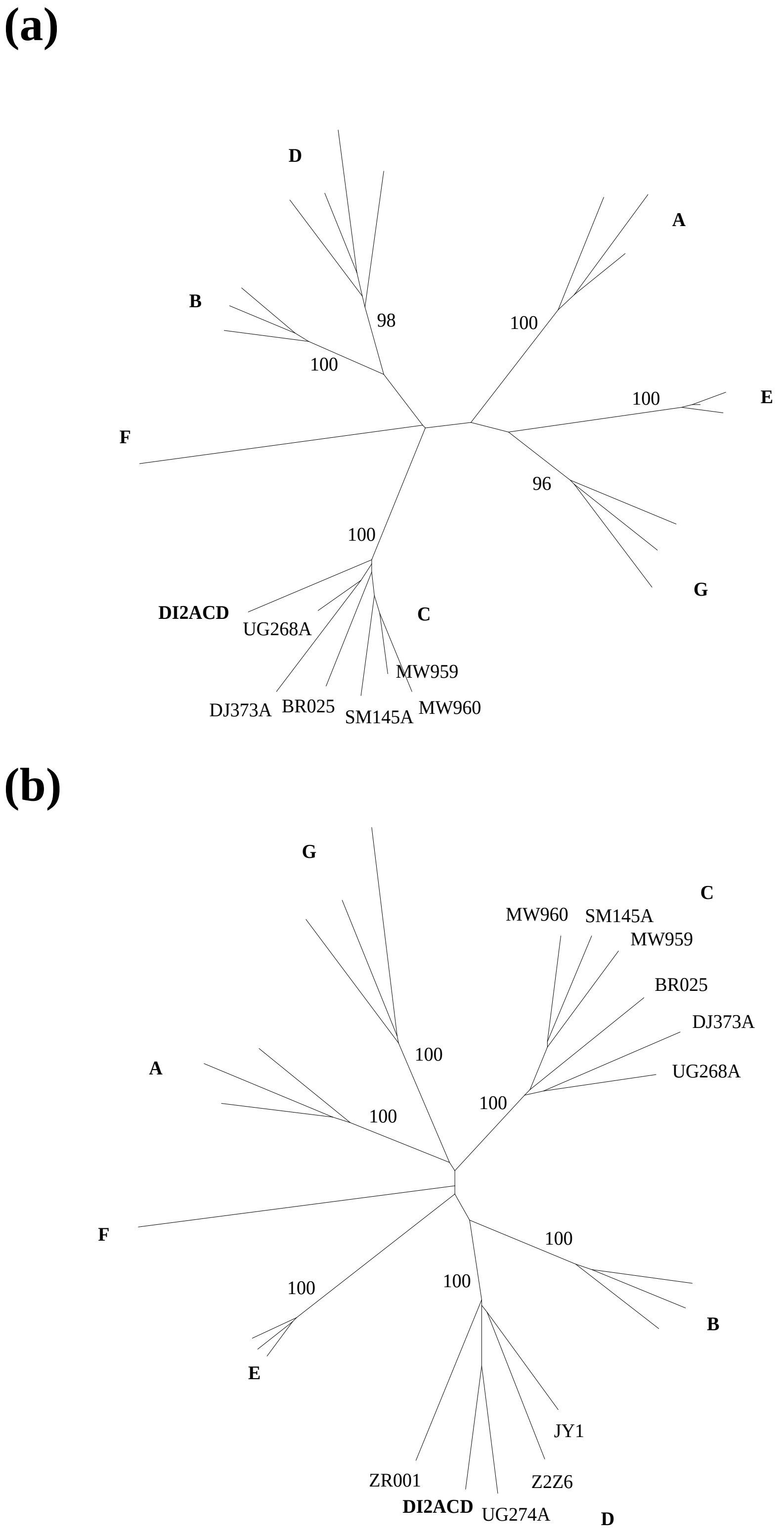

Figure 3. Phylogenetic trees showing two segments of sequence DI2ACD, an apparent C-D recombinant in *env*. Note that DI2ACD groups with the C clade in (a) (bases 1–324, 318 sites retained) and with the D clade in (b) (bases 325–1458, 1002 sites retained).

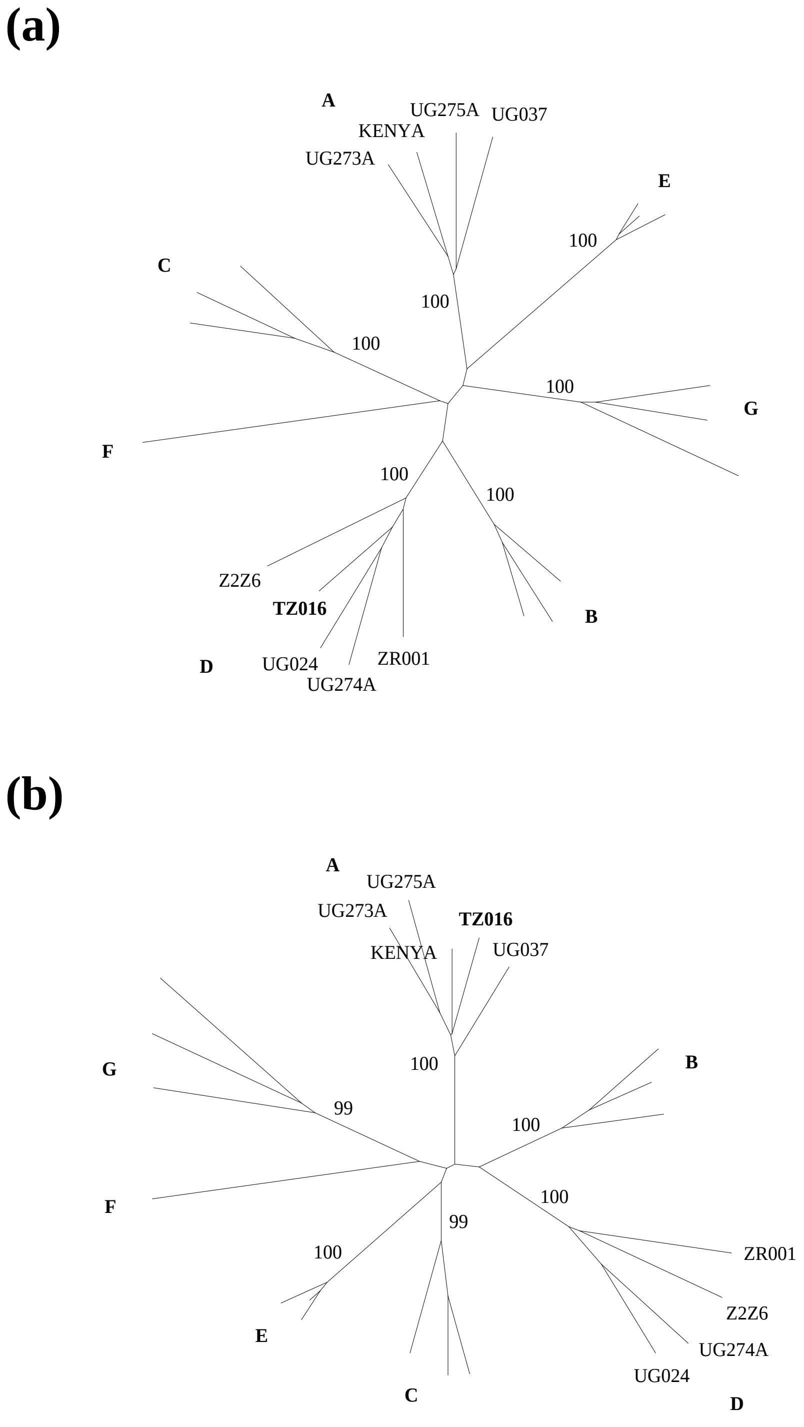

Figure 4. Phylogenetic trees showing two segments of sequence TZ016, an apparent A-D recombinant in *env*. Note that TZ016 groups with the D clade in (a) (bases 1–735, 666 sites retained) and with the A clade in (b) (bases 736–1113, 360 sites retained).

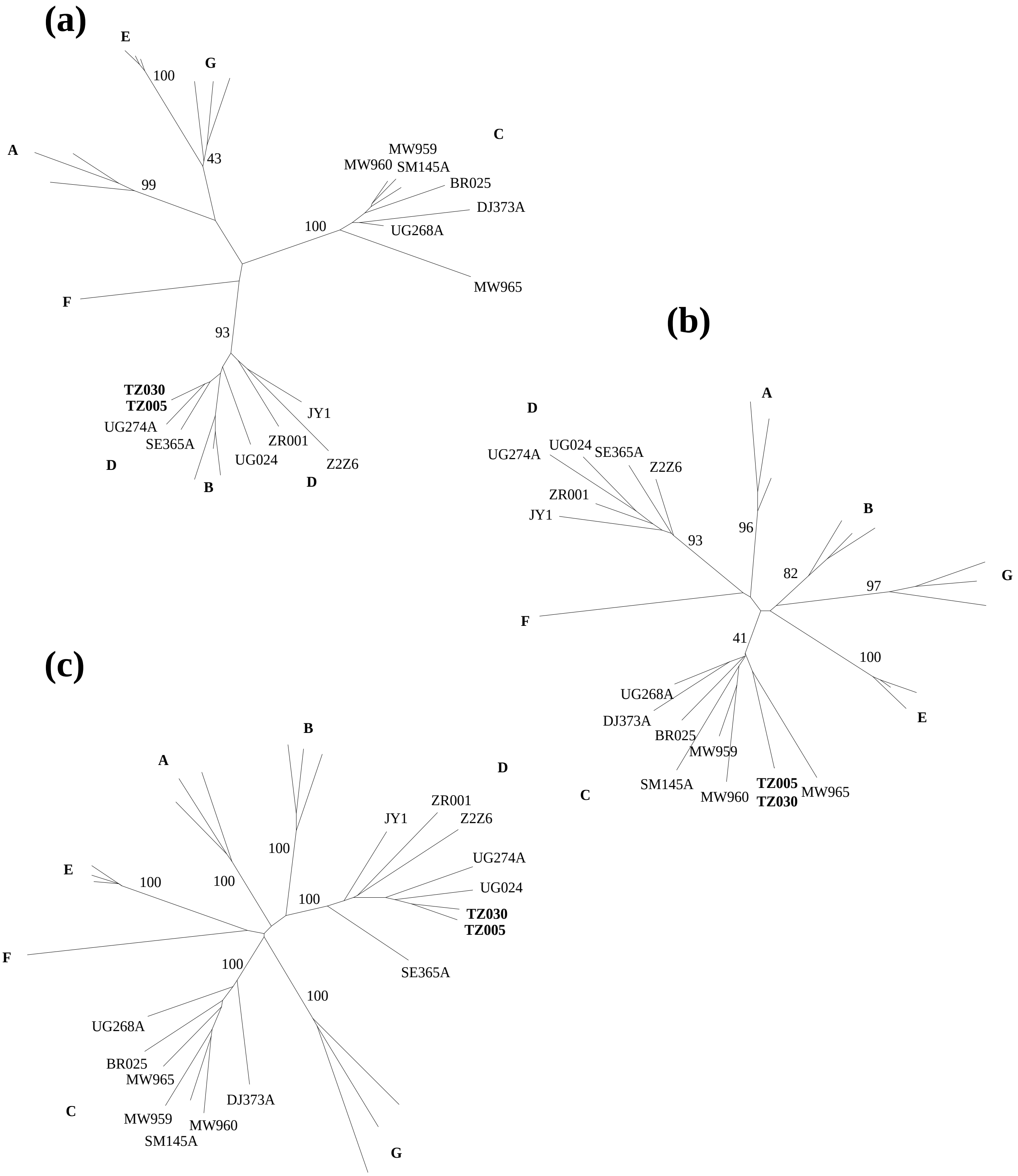

Figure 5. Phylogenetic trees showing three segments of *env* sequences TZ005 and TZ030. Although these sequences have been classified as D's [10], they are similar to sequences of the C clade in the segment shown in (b). Tree (a) includes bases 1–157 (144 sites retained), tree (b) includes bases 158–360 (203 sites retained), and tree (c) includes bases 361–1113 (679 sites retained).

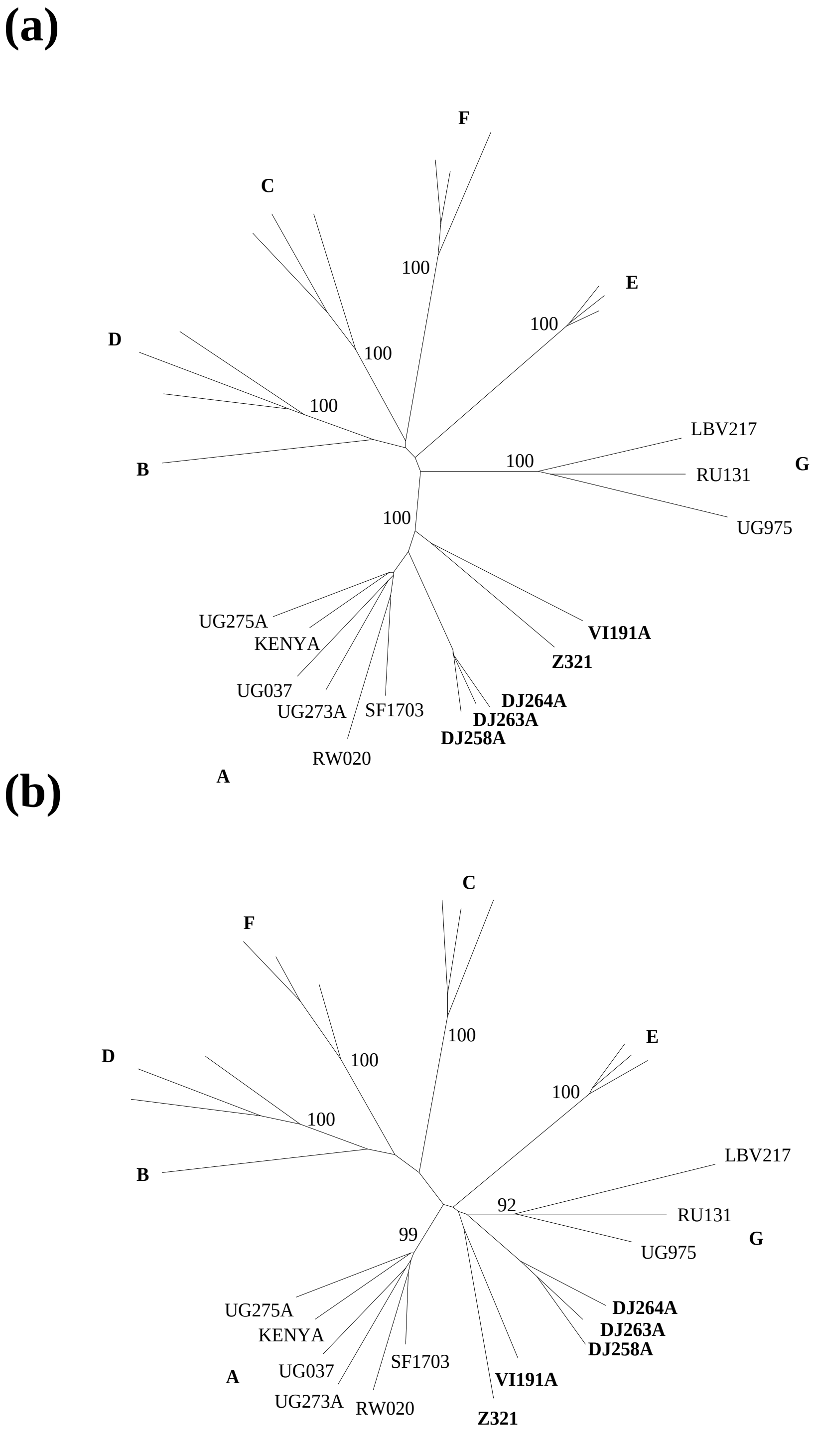

Figure 6. Phylogenetic trees showing two segments of sequences DJ258, DJ263, DJ264, VI191A, and Z321 in *env*. Note that all five sequences are outliers to the A clade in (a) (bases 1–1886 according to sequence DJ258, 1721 sites retained), but are nearer the G clade in (b) (bases 1887–2574, 634 sites retained).

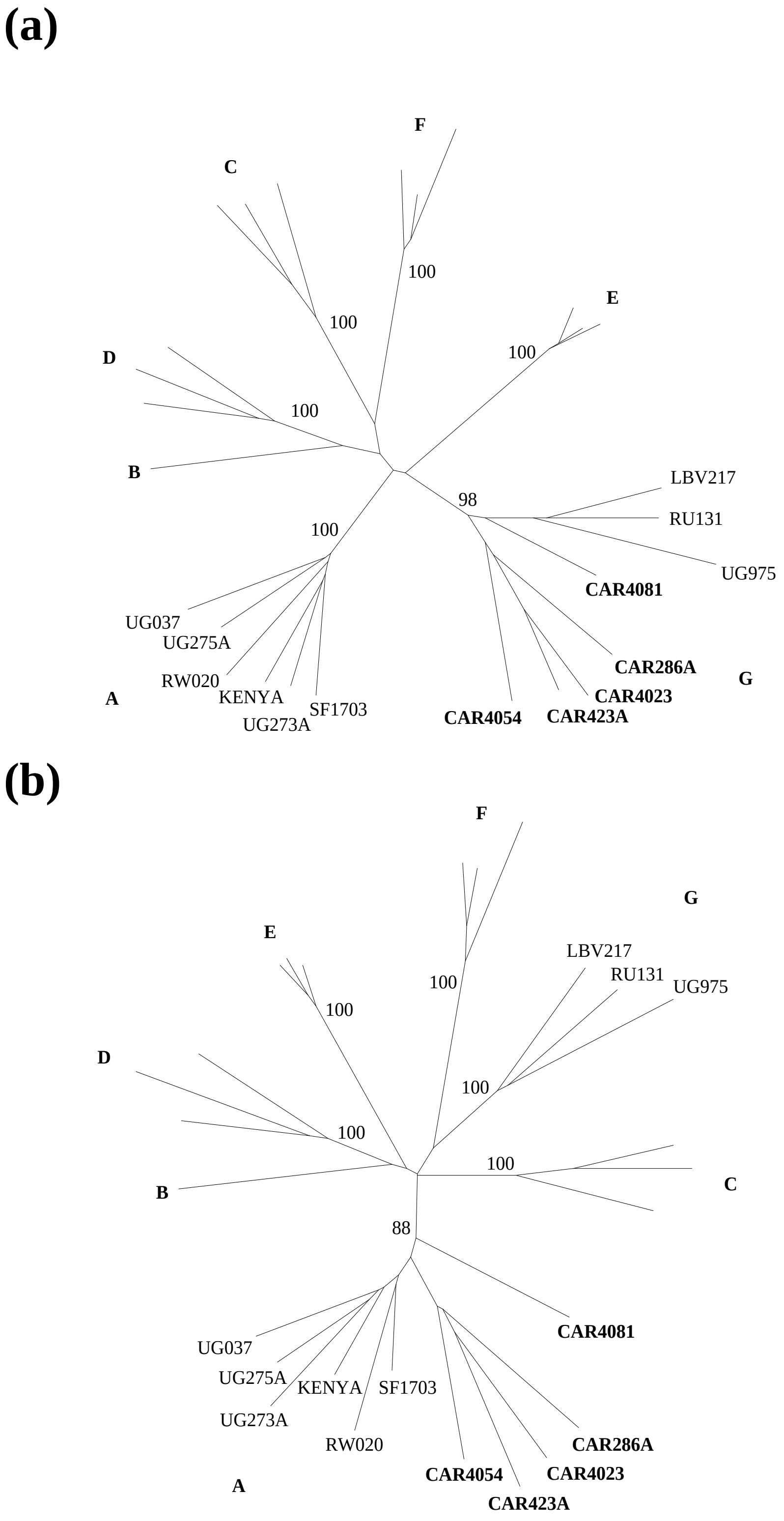

Figure 7. Phylogenetic trees showing two segments of sequences CAR286A, CAR4023, CAR4054, CAR4081, and CAR423A in *env*. Note that all five sequences are outliers to the G clade in (a) (bases 1–746 according to sequence CAR286A, 592 sites retained) and to the A clade in (b) (bases 747–1549, 710 sites retained).

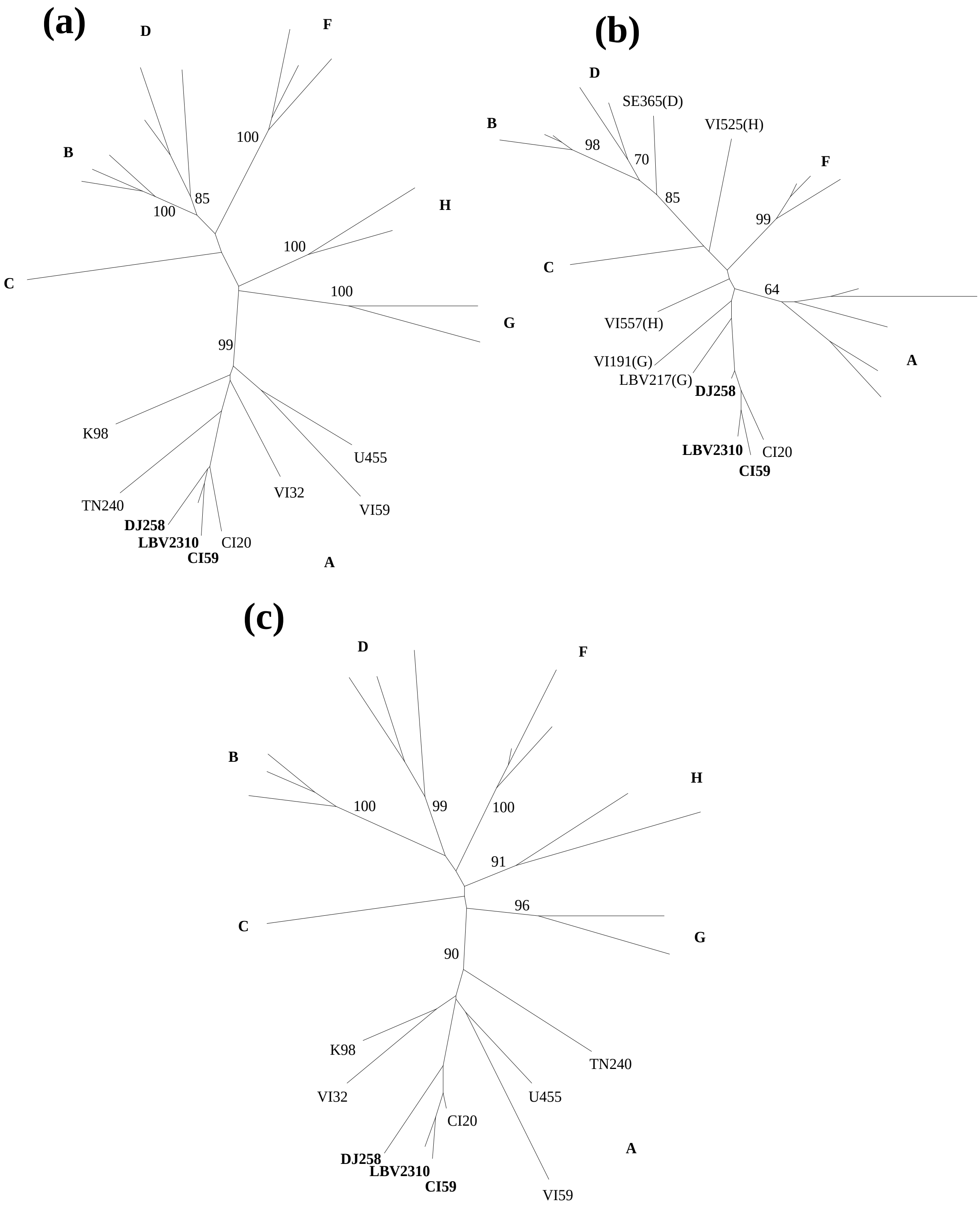

Figure 8. Phylogenetic trees showing three segments of sequences CI59, DJ258, and LBV2310 in *gag*. Note that these three sequences, along with the closely-related CI20, group with the A clade in (a) and (c), but that the A, G, and H clades break down in (b). Tree (a) includes bases 1–741 (715 sites retained), tree (b) includes bases 742–970 (224 sites retained), and tree (c) includes bases 971–1458 (440 sites retained). Positions in all three sequences are numbered the same because they are the same length and exactly homologous.